Patterns of Markup use in Wikipedia

1st Jonathan Martin
LIFIA, Facultad de Informática, UNLP
La Plata, Argentina
Jonathan.Martin@lifia.info.unlp.edu.ar

2nd Diego Torres
LIFIA, Facultad de Informática, UNLP
CIC, Prov. de Bs. As.
Departamento de Ciencia y Tecnología, UNQ
La Plata, Argentina
Diego.Torres@lifia.info.unlp.edu.ar

3rd Alejandro Fernandez
LIFIA, Facultad de Informática, UNLP
CIC, Prov. de Bs. As.
La Plata, Argentina
Alejandro.Fernandez@lifia.info.unlp.edu.ar

***Abstract*—Wikipedia is a knowledge building community that lets anyone create and edit articles. While editing articles, users employ visual structure elements (VSE) to format content. VSEs are part of the Wikipedia markup language. All creation and editing events are recorded in a revision history. An unsupervised learning approach was used to analyze a dataset with more than 2,000,000 revisions of 126,000 articles. Using K-Means clustering and association rules mining a general classification of revisions was derived. Relevant classes include vandalism revisions, correction revisions and common revisions. Each class was later studied, and patterns of usage of markups elements identified. Those results help to identify the user intention, and the knowledge of VSE use could contribute to improving the actual text editors provide by Wikipedia to improve the editor's activity finally.**

***Index Terms*—Pattern mining, Machine learning, Unsupervised learning, Wikipedia.**

I. Introduction

The knowledge building process emphasizes on the production and continuous improvement of knowledge pieces. In other words, it is a collective creation of public knowledge [1]. Some communities give support to this activity using a Web-based approach called knowledge building communities [2].

Wikipedia is a well-known example of these communities. In Wikipedia, any person can create or edit articles collaboratively by means a text editor provided. Currently, two version of editors could be used in Wikipedia: the one with wiki text (markup language), or a WYSIWYG editor. Both editors allow users to format the article content by using several visual structure elements (VSE) like links, headings, or lists, among others.

Every time a person saves a new edition in the content of an article, a revision with those changes is stored in its "revision history". A revision history includes the changes (revisions) that were made in chronological order of an article. For each revision in the revision history, it could be seen the author identification, date of creation, and the content of the complete article. Despite the selected Wikipedia editor, the content of each revision article includes VSE information.

978-1-5386-3483-7/17/$31.00 2017 IEEE

Revision history or log analysis is a research area to understand the evolution and the behavior over time. In the context of Wikipedia, the evolution of each article content is described in its revision history. With this information, it is possible, for example, to reconstruct the writing process from the beginning to the end [3]. It is also possible to detect and classify the user's behavior [4], to study the quality evolution of an article [5] or, to detect the use patterns of provided tools to perform the knowledge building process such as discussion pages, communication board, or format elements.

Several works are based on the study of wikis and their changes in different topics: study of editor's behavior to detect and classify editors into roles profile [6]–[8], automatic vandalism detection of vandal behavior and how the systems to detect vandalism works [9]–[11], analysis of activity level in the wikis along the time [12], [13], and semantical study or annotation of Wikipedia [14]–[16]. However, at the moment of writing this article, there are not evidence of approaches that explodes the VSE information to study the article content.

In this article, an unsupervised learning approach is introduced in order to apply cluster analysis and association rule mining to analyze VSE elements in the article content and in the article content evolution. The approach is structured in several steps. Firstly, cluster analysis is applied to study the use of VSEs in a revision and, among revisions. Secondly, association mining rules is used to make a co-occurrence analysis among revisions. Finally, a combination of both analysis is purposed to have a better detail in the pattern VSE use in Wikipedia editions. The study was conducted analyzing 126,000 of articles of the Wikipedia's English version that contain more than 2,000,000 of revisions. As a result, a classification for revisions was obtained, and the existence of VSE used patterns was detected. Those results can be useful to distinguish user's intentions or improve the article editors with the VSE patterns.

The remainder of this paper is as follows: Section II provides related works. Sections III presents the approach and the tools to work. Section IV describes the information of revision history, presents the dataset and details the evaluation results. Finally, Section V draws some conclusions and points

in a few directions for future investigation.

II. Related Work

In the context of Wikipedia revision history studies, Viegas et al. [17] introduce *history flow*, a tool to detect collaboration patterns: vandalism and repair, anonymity versus named authorship, negotiation, and content stability. Kiesel et al. [13] introduce an analysis spatio-temporal to detect vandal patterns. For that, they use the revision history, more specifically the reverse editions. All of these works are centered in authors behavior pattern detection. Though our approach also includes vandalism detection, it is centered in article content evolution.

Zeng et al. [18] analyze the changes in the history revision at the sentence level to assign a level of trust to each fragment. This level of trust is based on the author that create or modify a sentence. Javanmardi et al. [5], [19] present a model of the evolution of the articles based on their content quality. According to the authors, the quality of an article changes along the time among different revisions. In comparison, although our approach is not directly related to quality, the use VSE-quality-based analysis could be immersed in further work.

Edit activity patterns are studied in [20] by means of applying hierarchical cluster analysis to analyze time series of activity. In that work, six wiki edit activities were detected. In the works of Yang et al. [4], [6], different techniques to find user's roles and how they affect to the article quality are introduced. They apply Latent Dirichlet Allocation (LDA) unsupervised learning approach to discover the user's roles. In our work, unsupervised learning is specifically used in pattern discovering in VSE. Additionally, association rules mining and cluster analysis are used as well, but instead of a hierarchical algorithm, a K-means technique was used.

III. Methodology

This section introduces the methodology and the approach of this article. The methodology was guided by the following aspects:

1) Is it possible to group revisions by the analysis of theirs VSEs? It is desirable to recognize the presence of patterns in the use of VSE among revisions. For example, revisions from different articles that use the same set of VSEs.
2) From the changes of VSE among revisions (deletions and additions), that we call this changes "revision evolution". It is wanted to analyze and characterize this revision evolution. Is it possible to group different revision evolution's? Indeed, it would be desirable to perform this analysis with different levels of granularity, for example, distinguishing among low, medium or high level of additions or deletions.
3) What and how is the co-occurrence of those changes? For example, there are two VSEs called X and Y, when are X and Y applied together, which are their support, confidence or lift?

To analyze the former aspects, the tools and metrics for the study are presented below.

A. Tools

In this work, unsupervised learning was used in R software environment for statistical computing[2].Unsupervised learning is used in the exploratory analysis to search undiscovered patterns in not label data. To answering the questions about the existence of patterns of VSE use, K-means algorithm was used to cluster analysis; it was selected because It is fast with big datasets and allow to a straightforward interpretation. [21]. To answer questions related to the correlations of VSE, association rules mining algorithms were used. More specifically Eclat and Apriori to get support, confidence and lift for each rule or co-occurrence. Those algorithms were selected because both are well-known and fully documented allowing to a straightforward association rule mining for a frequent item set. As both algorithms are designed to work with transactions, in this work, a transaction is going to be considered equals to a revision. Because continuous data about the VSE applications or deletions is conserved in the revision evolution data, it was aggregated using different values intervals that there are specified for each transaction used. Finally, if well those algorithm are known for being stateless representation of data, this work studied the article content evolution using the differences among the revisions how is seen in the case studies section.

B. Metrics

The algorithms previously mentioned use some metrics in their applications. K-means require an assignation of the k number of clusters searched. *Calinski-Harabasz Index* [22]and *Average silhouette width* [23] metrics were applied to get a recommended cluster number. Also, the coefficient of Jaccard was employed to evaluate the stability of clusters. Jaccard coefficient give a measure of similarity between two clusters, iterating applying clustering and Jaccard coefficient the cluster stability of each cluster in the original clustering is the mean value of its Jaccard coefficient over all the iterations. [21]

Eclat and Apriori compute the support, confidence and lift. The support of an element (a set of co-occurrence elements in Eclat and a rule in Apriori) is the number of transactions that contain the element divided by the total number of transactions. In Apriori, the rules are like "if X, then Y". It means that every time the itemset X is seen in

[2]All code and dataset used is in: https://github.com/jonx18/Patterns-of-Markup-use-in-Wikipedias-Context

a transaction, see Y is also expected. The confidence can be represented such as $support(union(X,Y))/support(X)$, where the $union(X,Y)$ means that you are referencing to the rule that contains both X and Y. Thus, the confidence of the rule represents how often appears Y when X is there. Finally, lift compares the probability of an observed pattern with the probability that observe that pattern just by chance. The lift of a rule is given by $support(union(X,Y))/(support(X) * support(Y))$. If lift is 1, then X and Y are independent.

IV. Study Case

This section explains the structure and the different information can be obtained from the revision history of each article in Wikipedia. After, the dataset used to the study cases is presented, and finally, each study case it is presented with it's results.

A. Revision History

1) General revision information: Each article has a revision history which includes all the previous revisions ordered since earliest to the oldest. When an editor modifies an article and saves it, a new revision is created and put in the first position in the revision history.

The revision history has the article information like the title of the article (title), the article identification (id), the namespace (ns) to type the article (regular article, category, and others), and the revisions (revision). Each revision in the history has an identifier (id), time stamp (timestamp), the person who made it (contributor), comment (comment), and finally, the full article content at the moment of the current revision was created (text). Article's revision history is available in several formats, this work uses the XML format, and an example is shown in Listing 1.

```
<page>
  <title>Pope</title>
  <ns>0</ns>
  <id>23056</id>
  <revision>
    <id>2806055</id>
          ...
  </revision>
  <revision>
    <id>2806196</id>
    <parentid>2806055</parentid>
    <timestamp>2004-03-17T16:46:39Z</timesta
    <contributor>
      <username>Barbara Shack</username>
      <id>40231</id>
    </contributor>
    <text>{{About|the leader of the
          Catholic Church|the popes of
      other churches, and other uses}}...
    </text>
  </revision>
  <revision>
...
    </revision>
</page>
```

Listing 1. Article XML file example.

2) VSE in revision content: The revision content described with the XML tag `text` is a plain text within the Wikipedia markups. From these markups, the VSEs are extracted. For example, Figure 1 shows the editor version in plain text with the Wikipedia markups for the Pope article[3]. In the image, it can be seen the text ==*History*== which represents a heading level 2 with title *History* (Figure 2 shows the user view). The VSE that represents this markup will be the `heading2`.

A full list of VSEs analyzed in the current work is detailed in Table I. The first column details the VSE name, second column details the Markup name, third and fourth column detail the opening and ending tag of each Wikipedia markup respectively, and the description appears in the last column.

B. Dataset

The Dataset was composed for 2,000,000 revisions from 126,000 articles extracted from the English Wikipedia; only articles with name space 0 were used, in other words, common articles, not categories or talk pages. Those articles contained the revisions since their creation until December 28th, 2016 when they were extracted using Wikipedia API.

C. 1st case: Revisions Group Analysis

This case tried to describe how are grouped the revision by their VSEs using K-means to made clustering. As it

[3]https://en.wikipedia.org/wiki/Pope accessed on December 28th, 2016
[3]List of all articles used in https://rawgit.com/jonx18/Patterns-of-Markup-use-in-Wikipedias-Context/master/articlelist.html .

```
==History==
{{Main article|History of the papacy}}
===Title and etymology===
The word ''pope'' derives from [[Greek language|Greek]]
πάππας meaning "father". In the early centuries of
[[Christianity]], this title was applied, especially in the
east, to all bishops<ref name=ODCC:Pope></ref> and other
senior clergy, and later became reserved in the west to the
Bishop of Rome, a reservation made official only in the 11th
century.<ref></ref><ref></ref><ref></ref><ref></ref><ref></
ref> The earliest record of the use of this title was in
regard to the by then deceased [[Patriarch of Alexandria]],
[[Pope Heraclas of Alexandria]] (232–248).<ref></ref> The
earliest recorded use of the title "pope" in English dates
to the mid-10th century, when it was used in reference to
[[Pope Vitalian]] in an Old English translation of
[[Bede]]'s ''[[Historia ecclesiastica gentis
Anglorum]]''.<ref></ref>
```

Fig. 1. Extract from Pope's article in plain text.

TABLE I
ANALYZED VSES

VSE Name	Markup Name	Markup opening	Markup ending	Description
nowiki	Nowiki	<nowiki>	</nowiki>	Prevent the application of wiki-markups.
big	Big text	<big>	</big>	Increase the text size.
small	Small text	<small>	</small>	Decrease the text size.
sup	Superscripts	[	]	Create a superscripts.
sub	Subscripts	_		Create a subscripts.
s	Strike-through	<s>	</s>	strike out text.
blockquote	Blockquote	<blockquote>	</blockquote>	Create a blockquote.
includeonly	Includeonly	<includeonly>	</includeonly>	It is used for templates.
reference	Reference	<ref>	</ref>	Create a reference.
heading2	Heading 2	==	==	Create a title type 2.
heading3	Heading 3	===	===	Create a title type 3.
heading4	Heading 4	====	====	Create a title type 4.
heading5	Heading 5	=====	=====	Create a title type 5.
italic	Italic text	Two apostrophes	Two apostrophes	Italicize text.
blod	Bold text	Three apostrophes	Three apostrophes	Bold the text.
italicblod	Italic and bold text	Five apostrophes	Five apostrophes	Italic and bold formatting.
external	External links	[	]	Links to web pages outside Wikipedia.
internal	Interwiki link	[[	]]	Linking to a page on another wiki in English.
numberedelement	Ordered Element	#	Not need	Create ordered element.
bulletedelement	Unordered Element	*	Not need	Create unordered element.
redirect	Redirects	#REDIRECT [[	]]	Redirect one article title to another.
indent2	Indent text 2	Two colon ::	Not need	Create indent text type 2.
indent1	Indent text 1	Colon :	Not need	Create indent text type 1.
infobox	InfoBox	{{Infobox	}}	Create an Infobox.
wikitable	WikiTable	{|class=wikitable	|}	Make a table.
cite	Cite	{{cite	}}	Create a cite.

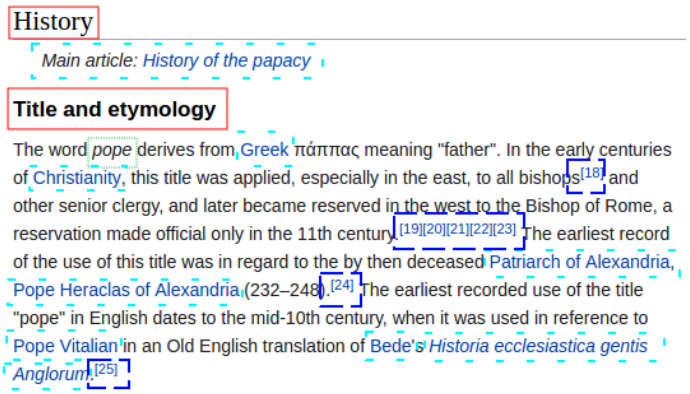

Fig. 2. Extract from Pope's article in Wikipedia.

took more than a day, with the full dataset, the search was evaluated in an iterative strategy. It consisted in evaluate K-means over the revisions of a number P articles, and finally, increasing P by one. Table II shows the results of three iterations. The first column shows the number of articles used, and the second and third column shows the number of clusters recommended by the criterion *Calinski-Harabasz Index* (CH) and *Average silhouette width*(ASW) respectively. In the first row, the recommendations were 17 and 11, but in the second row, those recommendations began to be greater with values as 18 in both cases. Finally, in the third row, the values are increased more.

The constant increment of the number of recommended clusters shown that the articles created clusters to themselves. It was because the visual structure(composed by VSEs) among revisions of the same article tends to be similar.

TABLE II
CLUSTERS RECOMMENDED.

Number of Articles	CH	ASW
3	17	11
4	18	18
5	20	20

D. 2nd case: Revision Evolution Analysis

In this case, groups into the revision evolution were searched. The quantitative changes of VSE between revisions of same articles were used. For example, between two revisions of the same article a VSE heading was added, and the addition was kept as information of the revision to compute the clustering with K-means. K-means was run with 3 clusters by recommendation of the metrics CH and ASW.The clusters are represented in Figure 3. The cluster 3 (in the center of the figure) is the one with more elements which are grouped in a more cohesive manner than the other two clusters. Moreover, the other two clusters are in opposite positions where both of them have more disperse elements.

After, a subset of revisions from each cluster were hand analyzed regarding a typological of changes. What was the comment of the revision? Was the revision a revert? From this analysis and considering the descriptive statistics of each cluster , three types of revisions could be determined, and each was named based on its characteristics:

- **Vandal Editions:** Those are revisions with a high decrement in a VSE. Also, those revisions represented dele-

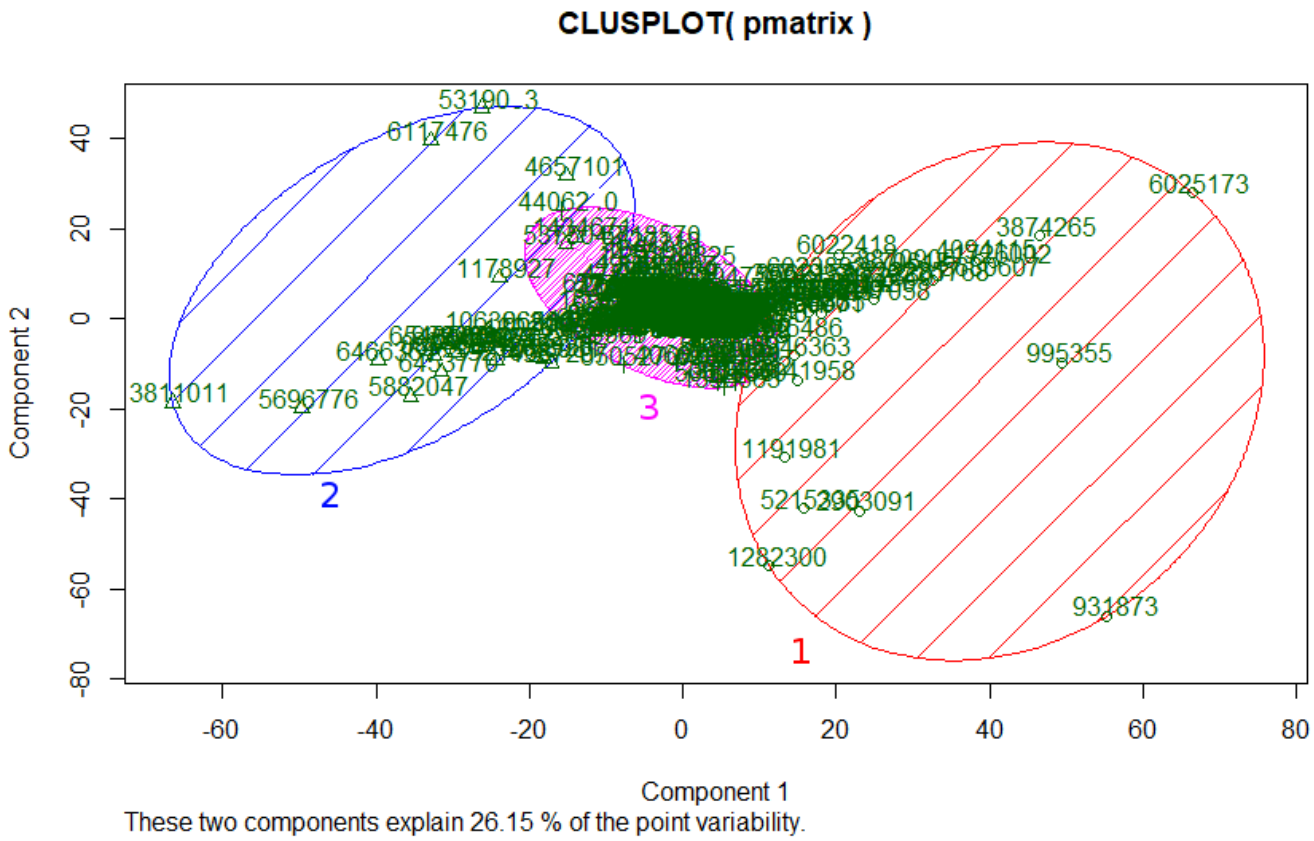

Fig. 3. K-means clusters.

tions or replacement of content with damaging intentions to the article. It was represented by the cluster 1.

- **Correction Editions:** Those are revisions that recover the content of Vandal Editions. Consequently, it have a high increment of VSE. Also, many of those revision had comments explaining those are reverts or vandal correction. It was represented by the cluster 2.
- **Common Editions:** Those are revisions with changes in a normal range for revisions in articles. Also, those revisions usually had detailed comments of the changes or at most, marked as self-revert. It was represented by the cluster 3.

E. 3rd case: Addition and Deletion Patterns

This case analyzed the existence of VSEs addition or deletion patterns. To analyze the revision evolution in this way, the VSE information was aggregated and tagged by the following next rules:

- When any deletion of VSE was detected, the tag *"-Deleted"* was applied.
- When any addition of VSE was detected, the tag *"-Added"* was applied.
- If changes were not detected, the VSE was deleted to prevent bias the results.

In order to perform this analysis, Eclat was configured with a minimum support of 0.01 and a least of 2 elements. Also, Apriori was configured with support and confidence minimum or 0.01 and a least of 2 elements in the first part of the rule. Those configurations helped to focus in the relations between VSE.

As result[4], the Table IV for Eclat was obtained. First column details the identification of the results, the second column shows the VSE that co-occur and the support is in the last column. There were items correlated with a certain support, e.g., the first item of the table means that the VSE's that represents a blockquote and the indicator of templates inclusions are added together with 3.39% of support.

The results for Apriori are shown in Table III. First column there is the identifier of the rule, the second column the rule in the form of "If X then Y" is presented, and the remaining columns represent the metrics support, confidence and lift respectively. The Apriori's rules could be read such as "If some VSE was added/deleted then

[4]The full results Tables can be seen in: https://github.com/jonx18/Patterns-of-Markup-use-in-Wikipedias-Context/blob/master/3-Thirdcase/3-Thirdcaseresults.xlsx

some other VSE was added/deleted", and the rule was evaluated with support, confidence and lift. Those results were ordered by confidence from highest to lowest. The first seven rules had a confidence over 50% and all with lift over 1.

The results showed that those rules and the co-occurrence were only between the same events, in other words, co-occurrence on VSEs were additions or deletions but not crosses of them. As a result, the existence a correlation in the way of VSEs were used together can be affirmed but was necessarily looked deep to confirm that, and It was done in the next case.

F. 4th case: Fine granularity of changes

In this case, the changes on the revision evolution in a fine granularity were analyzed. The VSE information was aggregated and tagged with three quantity ranges called Low, Medium, and High. The following rules were used:

- When any deletion of VSE was detected, the tags *"-Low-","-Medium-","-High-"* were used in the range of VSE deleted was between (0 : -2], (-2 : -10] and (-10:-Inf).
- When any application of VSE was detected, the tags *"+Low+","+Medium+", "+High+"* were used in the range of VSE applied was between (0 : 2], (2 : 10] and (10:Inf).
- If changes were not detected, the VSE was deleted to prevent bias the results.

In order to perform this analysis, Eclat was configured with a minimum support of 0.001 and a least of 2 elements. Also, Apriori was configured with support and confidence minimum or 0.001 and a least of 2 elements in the first part of the rule. Those configurations helped to focus in the relations between VSE and the frequency of their changes properly. Also, the minimums selected allow getting more interesting results in a metric of lift refer.

As result[5], the Table VI for Eclat was obtained. First column details the identification of the results, the second column shows the VSE that co-occur and the support is in the last column. There were items correlated with a certain support, e.g., the first item of the table means that the VSE's to represent block-quote and the indicator of templates inclusions are added together at a low level with 2.15% of support. The items obtained from those results had less support than the results of the third case.

The results for Apriori are shown in Table V. First column there is the identifier of the rule, the second column the rule in the form of "If X then Y" is presented, and the remaining columns represent the metrics support, confidence and lift respectively. Rules can be read such as "If some VSE was added/deleted in a level some other VSE would be added/deleted in another level" and the rule was evaluated with support, confidence. From Apriori 298 rules were obtained, those rules had a higher confidence than the results in the third case; also, the lift was also high. Finally, all of the first results or the results with the highest confidence were related to events of large deletions or significant additions, but those rules had a shallow support.

From those results, the existence of a direct correlation between the application or deletion of VSE together could be confirmed. There was less support than the support in the second case, but that was because the events of deletion and addition were split into three small events each. Besides this in Apriori's results, all results were of high value, it was because the little support allows Vandal or Correction revisions and those values biased the rules. This bias into the results was corrected in the implementation of the next case.

G. 5th case: Fine Revision Evolution by Cluster

This case analyze the characteristics of the clusters (Vandal, Correction and Common editions) found in the second case but using the process of the fourth case.

First, to get a more refined version of the clusters, an iterative clustering process was done. In each iteration, the clusters were evaluated with Jaccard coefficient, and VSE elements of the clusters with less stability were deleted. This iteration continues until the coefficient of the clusters did not increase more. Those clusters were kept, and the deleted elements were assigned to the stabilized clusters that correspond.

After clustering, for each cluster that corresponds with one of three types of revision described in the second case, the revisions were analyzed applying the process described in the fourth case. In Figures 4 and 5, value's ranges for each metric in Eclat and Apriori for each cluster are shown. Regarding the support, the clusters of Vandal and Correction Editions had the highest support, and that was because those clusters had fewer transactions and with particular items. In another hand, the support of the Common Editions was the lowest, but it was expected because there were more variety of VSEs added or deleted in different magnitudes and this cluster was the biggest. On confidence, also the clusters of Vandal and Correction Editions had the highest confidence. Those levels of confidence were because the confidence is sensitive to the frequency of the elements Y, and in this case, with elements with strong support higher confidence values were produced even if there existed no association between the items [24]. The Common Editions had an acceptable max level of confidence. Finally, about lift, the clusters of Vandal and Correction Editions had the worse lift level indicating that those rules were not patterns, in contrast, the Common Editions had levels of lift indicating those rules

[5]The full results Tables can be seen in: https://github.com/jonx18/Patterns-of-Markup-use-in-Wikipedias-Context/blob/master/4-Fourthcase/4-Fourthcaseresults.xlsx

TABLE III
CASE 3 APRIORI OUTPUT.

Id	X =>Y	Support	Confidence	Lift
[1]	{s=+Added}=>{includeonly=+Added}	0.01112287	0.7554439	13.215852
[2]	{external=-Deleted}=>{indent1=-Deleted}	0.01038630	0.6004355	20.110261
[3]	{includeonly=+Added}=>{blockquote=+Added}	0.03391964	0.5933955	7.576177
[4]	{indent1=-Deleted}=>{internal=-Deleted}	0.01552891	0.5201064	12.117157
[5]	{includeonly=-Deleted}=>{blockquote=-Deleted}	0.01251420	0.5188473	16.603750
[6]	{external=+Added}=>{indent1=+Added}	0.01335142	0.5156500	9.520728
[7]	{indent1=+Added}=>{internal=+Added}	0.02732596	0.5045341	6.857714
[8]	{external=+Added}=>{bulletedelement=+Added}	0.01177999	0.4549593	5.379638
[...]	...	...	...	...
[20]	{blockquote=+Added}=>{bulletedelement=+Added}	0.01124470	0.1435667	1.697596

TABLE IV
CASE 3 ECLAT OUTPUT.

ID	VSE Co-Ocurrences	Support
[1]	{blockquote=+Added,includeonly=+Added}	0.03391964
[2]	{internal=+Added,indent1=+Added}	0.02732596
[3]	{internal=+Added,bulletedelement=+Added}	0.02166337
[4]	{internal=-Deleted,indent1=-Deleted}	0.01552891
[5]	{internal=-Deleted,bulletedelement=-Deleted}	0.01432017
[6]	{blockquote=+Added,big=+Added}	0.01347192
[7]	{external=+Added,indent1=+Added}	0.01335142
[8]	{bulletedelement=+Added,indent1=+Added}	0.01320869
[...]	....	...
[20]	{italic=+Added,bulletedelement=+Added}	0.01021900

such as patterns. Table VII has the rules mined [6] for the cluster 3 of Common Editions. In those rules can be seen the existence of a tendency to grow up in the articles, it is because the portion of rules with additions is bigger than the deletions.

The most significant assumptions that could be concluded from these results were:

- Internal links to Wikipedia are usually used with indent and bullets.
- It is common see that external links and references are applied and deleted together with indent.
- Also, cites and references are deleted together, but the additions of both VSEs together are occasional.
- Frequently, in the same revision that a heading of type 2 is added an indent of type 2 is added.
- The additions and the deletions hardly ever are performed together; only two rules were obtained that represent this activity and the most relevant was the interchange from an external link to an internal link.

V. CONCLUSIONS AND FURTHER WORK

This work presented an approach to studying the evolution of the articles in Wikipedia according to visual structural elements (VSE). The concept of VSE was introduced, and its importance was explained. Finally, an evaluation with more

[6]The full results Tables can be seen in: https://github.com/jonx18/Patterns-of-Markup-use-in-Wikipedias-Context/tree/master/5-Fifthcase

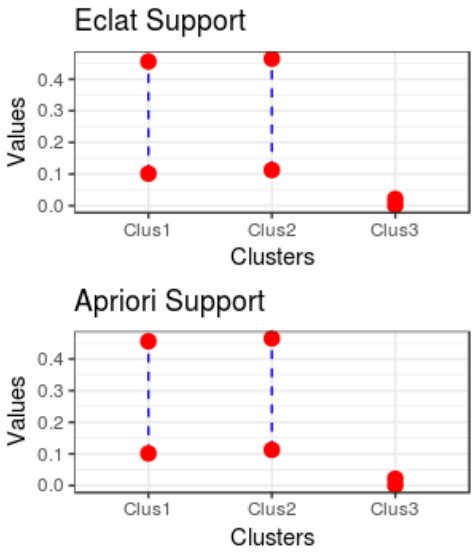

Fig. 4. Ranges of values in Eclat and Apriori for each cluster.

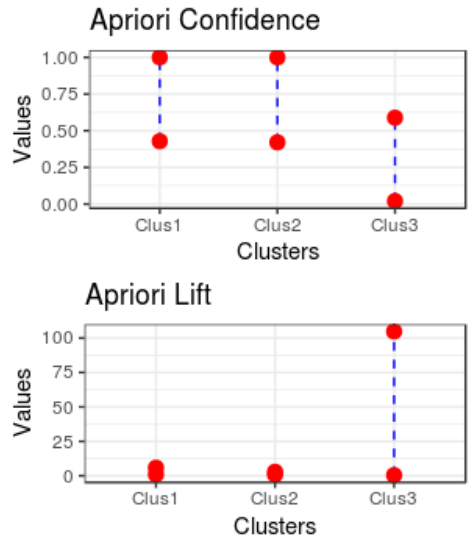

Fig. 5. Ranges of values in Eclat and Apriori for each cluster.

TABLE V
CASE 4 APRIORI OUTPUT.

Id	X =>Y	Support	Confidence	Lift
[1]	{italic=-High-}=>{internal=-High-}	0,0029	0,82068276	94,05664
[2]	{heading2=-High-}=>{bulletedelement=-High-}	0,00158	0,80456758	132,92482
[3]	{external=-High-}=>{indent1=-High-}	0,00192	0,79679373	150,12876
[4]	{indent1=-High-}=>{internal=-High-}	0,00422	0,79543123	91,16262
[5]	{external=-High-}=>{internal=-High-}	0,0019	0,78828856	90,34402
[6]	{external=-High-}=>{bulletedelement=-High-}	0,00189	0,7870648	130,03314
[7]	{italic=+High+}=>{internal=+High+}	0,00277	0,76948238	93,73189
[8]	{external=+High+}=>{internal=+High+}	0,00179	0,75399282	91,84508
[...]	...	...	...	...
[298]	{reference=+Low+}=>{italicblod=+Low+}	0,00104144	0,04053489	1,1357819

TABLE VI
CASE 4 ECLAT OUTPUT.

ID	VSE Co-Ocurrences	Support
[1]	{blockquote=+Low+,includeonly=+Low+}	0,02150432
[2]	{internal=+Low+,indent1=+Low+}	0,01552964
[3]	{internal=+Low+,bulletedelement=+Low+}	0,01300579
[4]	{heading2=+Low+,includeonly=+Low+}	0,00800723
[5]	{external=+Low+,indent1=+Low+}	0,00736968
[6]	{blockquote=+Low+,big=+Low+}	0,00697477
[7]	{s=+Low+,includeonly=+Low+}	0,00685411
[8]	{bulletedelement=+Low+,indent1=+Low+}	0,0068472
[...]	...	...
[460]	{italic=+Low+,heading2=+Low+}	0,001000242

than 2.000.000 revisions was described.

The evaluation was based on cluster analysis and association rules mining techniques. Of the evaluation can be concluded that using VSE as study element it is possible to classify the revisions in Vandal, Correction, and Common editions. Also, patterns about the application or deletion of VSE were found showing how are the VSEs used together and in what level they represent the real use of this facility to structure articles given by Wikipedia.

In future work, other techniques like LDA would like to be applied to VSE to improve and evaluate these results. Also, a classification algorithm would like to be created using supervised learning to classify the revisions in the classifications described here, because the unsupervised learning algorithms are more for exploratory analysis and they are not efficient for this kind of tasks. Finally, this work was applied only over an article subset of Wikipedia and if this study is applied over whole Wikipedia, the results might be more relevant or shown undiscovered VSE patterns. Moreover, this article focused specifically on Wikipedia, a general purpose knowledge building community. Future work will explore whether similar pattenrs are found in more specific Wikis such as those used by communities of practice in agriculture.

Acknowledgments

Authors of this publication acknowledge the contribution of the Project 691249, RUC-APS: Enhancing and implementing Knowledge based ICT solutions within high Risk and Uncertain Conditions for Agriculture Production Systems (www.ruc-aps.eu), funded by the European Union under their funding scheme H2020-MSCA-RISE-2015.

REFERENCES

[1] J. Moskaliuk, J. Kimmerle, and U. Cress, "Wiki-supported learning and knowledge building: effects of incongruity between knowledge and information," *Journal of Computer Assisted Learning*, vol. 25, pp. 549–561, 11 2009.

[2] R. G. Baraniuk, C. S. Burrus, D. H. Johnson, and D. L. Jones, "Sharing Knowledge and Building Communities in Signal Processing," no. September, 2004.

[3] O. Ferschke, "The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia," 7 2014.

[4] D. Yang, A. Halfaker, R. Kraut, and E. Hovy, "Edit Categories and Editor Role Identification in Wikipedia," 2013.

[5] S. Javanmardi and C. Lopes, "Statistical measure of quality in Wikipedia," *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pp. 132–138, 2010.

[6] D. Yang, A. Halfaker, R. Kraut, and E. Hovy, "Who Did What: Editor Role Identification in Wikipedia," 2016.

[7] T. H. McCormick, R. Ferrell, A. F. Karr, and P. B. Ryan, "Modeling User Reputation in Wikis," *Science And Technology*, vol. 4, no. 5, pp. 497–511, 2010.

[8] R. S. R. Geiger and A. Halfaker, "Using edit sessions to measure participation in Wikipedia," *Proceedings of the 2013 conference on ...*, no. February 2013, pp. 861–869, 2013.

[9] K. Smets, B. Goethals, and B. Verdonk, "Automatic Vandalism Detection in Wikipedia : Towards a Machine Learning Approach," *Proceedings of AAAI 2008 Workshop on Wikipedia and Artificial Intelligence An Evolving Synergy WikiAI08 (2008)*, pp. 43–48, 2008.

[10] A. Sarabadani, A. Halfaker, and D. Taraborelli, "Building automated vandalism detection tools for Wikidata," vol. 24, pp. 1647–1654, 2017.

[11] S. M. Mola Velasco, "Wikipedia vandalism detection through machine learning: Feature review and new proposals: Lab report for PAN at CLEF 2010," *CEUR Workshop Proceedings*, vol. 1176, 2010.

[12] A. Halfaker, R. S. Geiger, J. T. Morgan, and J. Riedl, "The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline," *American Behavioral Scientist*, vol. 57, no. 5, pp. 664–688, 2012.

[13] J. Kiesel, M. Potthast, M. Hagen, and B. Stein, "Spatio-temporal Analysis of Reverted Wikipedia Edits," no. 2010, 2017.

[14] R. Schenkel, F. M. Suchanek, and G. Kasneci, "YAWN: A Semantically Annotated Wikipedia XML Corpus," *Proceedings of GIFachtagung für Datenbanksysteme in Business Technologie und Web BTW2007*, vol. 103, pp. 277–291, 2007.

TABLE VII
CASE 5 APRIORI OUTPUT. COMMON EDITIONS

Id	X =>Y	Support	Confidence	Lift
[1]	{indent2=+Low+}=>{heading2=+Low+}	0,005381	0,588889	25,78025
[2]	{s=+Low+}=>{includeonly=+Low+}	0,007411	0,58871	13,06034
[3]	{blockquote=+High+}=>{includeonly=+Medium+}	0,002132	0,567568	68,17732
[4]	{s=-Medium-}=>{includeonly=-Medium-}	0,001421	0,518519	62,28546
[5]	{blod=-Medium-}=>{internal=-High-}	0,001117	0,5	104,7872
[6]	{blockquote=+High+}=>{italicblod=+Low+}	0,001827	0,486486	13,7698
[7]	{includeonly=+Low+}=>{blockquote=+Low+}	0,021421	0,475225	8,343972
[8]	{reference=-Medium-}=>{indent1=-Medium-}	0,002843	0,451613	39,71774
[...]	...	...	...	...
[215]	{internal=-Low-}=>{includeonly=-Low-}	0,001116751	0,02022059	0,4485874

TABLE VIII
CASE 5 ECLAT OUTPUT. COMMON EDITIONS

ID	VSE Co-Ocurrences	Support
[1]	{blockquote=+Low+,includeonly=+Low+}	0,021421
[2]	{internal=+Low+,indent1=+Low+}	0,016751
[3]	{internal=+Low+,bulletedelement=+Low+}	0,013096
[4]	{heading2=+Low+,includeonly=+Low+}	0,008122
[5]	{s=+Low+,includeonly=+Low+}	0,007411
[6]	{external=+Low+,indent1=+Low+}	0,00731
[7]	{reference=+Low+,indent1=+Low+}	0,00731
[8]	{bulletedelement=+Low+,indent1=+Low+}	0,007208
[...]	...	...
[251]	{blockquote=+Low+,bulletedelement= +Low+,includeonly=+Low+}	0,001015228

[15] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," *Proceedings of the National Conference on Artificial Intelligence*, vol. 21, no. 2, p. 1419, 2006.

[16] S. P. Ponzetto, S. P. Ponzetto, M. Strube, and M. Strube, "Exploiting semantic role lebeling, WordNet and Wikipedia for coreference resolution.," *proceedings of NAACL 2006*, vol. 33, no. June, pp. 192–199, 2006.

[17] F. B. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, vol. 6, no. 1, pp. 575–582, 2004.

[18] H. Zeng, R. Fikes, and K. Systems, "Mining Revision History to Assess Trustworthiness of Article Fragments," 2006.

[19] S. Javanmardi and Y. Ganjisaffar, "Statistical measure of the effectiveness of the open editing model of Wikipedia," *4th Int'l AAAI Conference*, 2010.

[20] O. Arazy and A. Croitoru, "The Sustainability of Corporate Wikis: A Time-Series Analysis of Activity Patterns," *ACM Transactions on Management Information Systems*, vol. 1, no. 1, pp. 1–24, 2010.

[21] N. ZUMEL, J. Mount, and J. Porzak, "Practical data science with R," p. 417, 2014.

[22] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[23] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[24] L. M. Sheikh, B. Tanveer, M. A. Hamdani, S. Mustafa, and A. Hamdani, "Interesting Measures for Mining Association Rules," *Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International*, pp. 641–644, 2004.